A Simple yet Scalable Granger Causal Structural Learning Approach for Topological Event Sequences

Mingjia Li*, Shuo Liu*, Hong Qian†, Aimin Zhou
(* Equal contribution † Corresponding author)

Shanghai Institute of AI for Education
School of Computer Science and Technology
East China Normal University, China

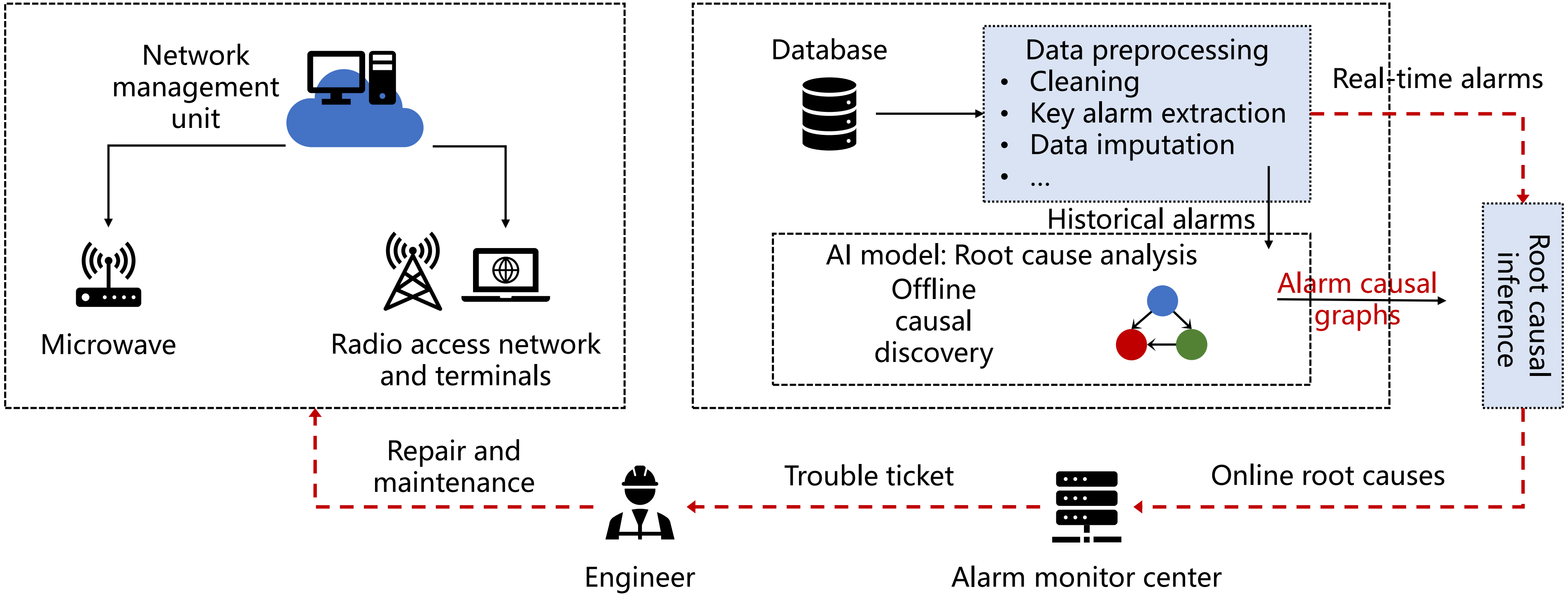
NeurIPS 2024

Outline

- Background
- Problem Formulation
- Challenges
- S^2GCSL
- Summary & Discussion

Background

RCA solution in TNFD



Goal: Telecommunication network fault diagnosis (TNFD).

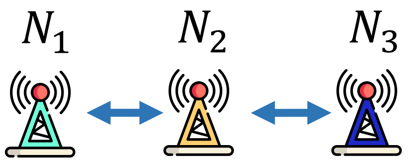
Method: Root cause analysis (RCA) is to learn a causal graph that represents alarm activation relations. and then using decision-making techniques to efficiently identify the **root cause alarm** when a fault occurs.

Problem: solve a **causal structure learning problem** AIOps (Artificial Intelligence for IT Operations).

Problem Formulation

An illustrative example of the topological event sequences generated by a telecommunication network

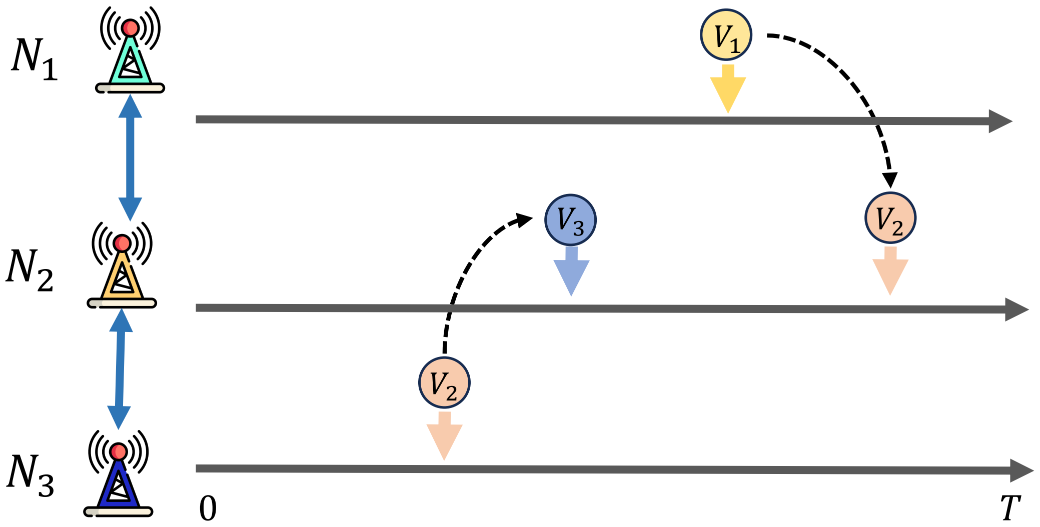
$\mathbf{B} \in \{0,1\}^{|N|\times|N|}$



Devices (Locations)

$\mathbf{A} \in \{0,1\}^{|V|\times|V|}$, $a\ DAG$

Alarms (Events)



Infer

$\boldsymbol{X} = \{(v_i, n_i, t_i) \mid i = 1, \dots, m\}$: Event sequence

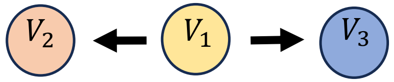
$v_i \in V$: Type of alarms (events)

$n_i \in N$: Devices

$t_i \in [0, T]$: Time domain

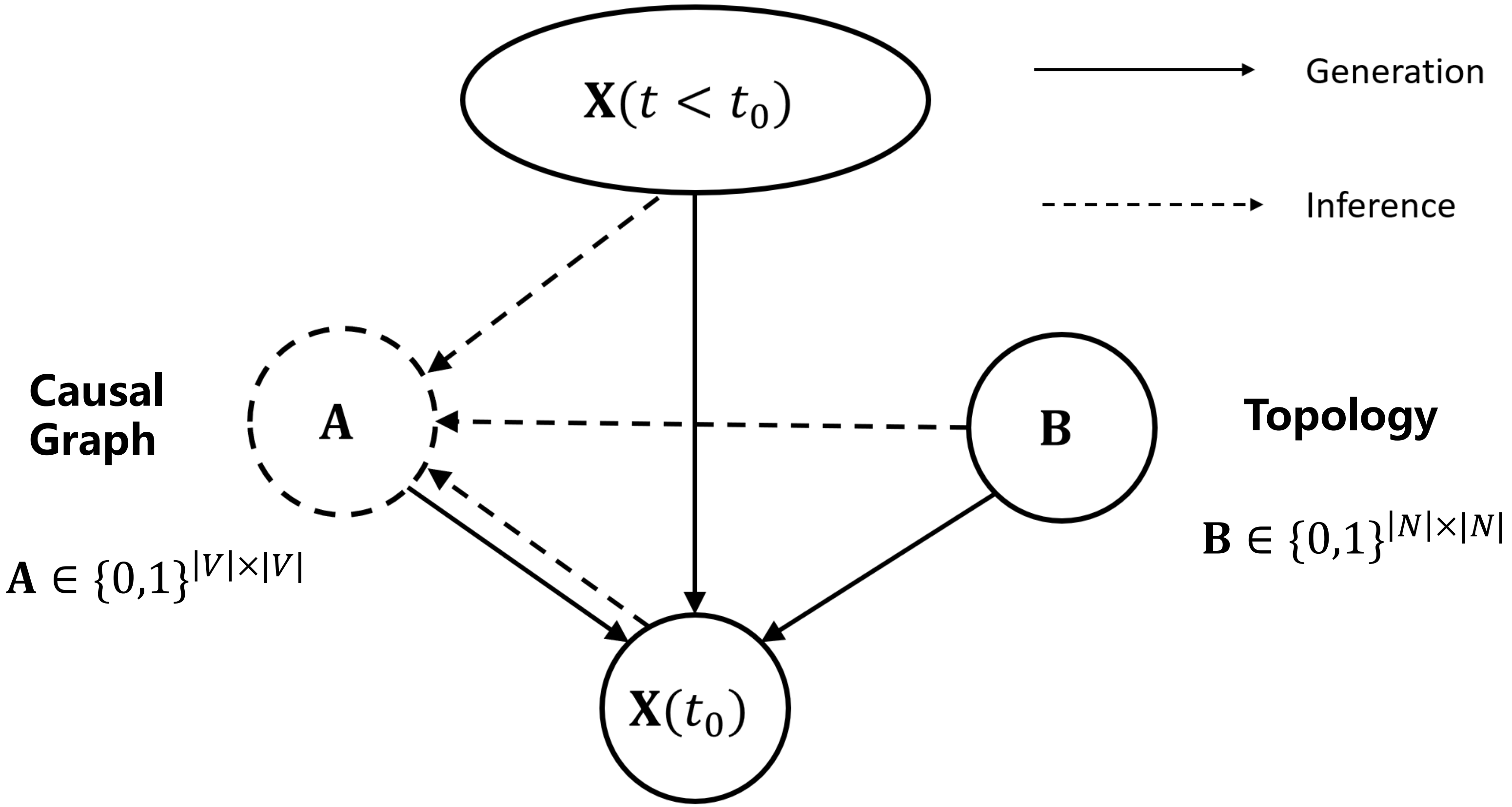
Root Cause Analysis (RCA)

Learned Causal Structure



Problem Formulation

Illustration of Data Generation and Causal Discovery Process in RCA



The solid lines represent the **data generation process.**

The dashed lines represent the **RCA inference process.**

Problem Formulation

Hawkes Process [1] $\lambda(t) = \mu + \sum_{t_i < t} \phi(t - t_i)$

- $\lambda(t)$ is the intensity function.
- μ is a **constant**, representing the **baseline intensity** of the event.
- The second term represents the **influence of events occurring before time t** on the intensity at time t, where ϕ is a decay function.

Topological Multivariate Hawkes Process [2] $\lambda_v^n(t) = \mu_v^n + \sum_{i:n_i \in Nei(n), t_i < t} a_{v_i v} \phi(t - t_i)$

- $\lambda_v^n(t)$ represents the **intensity function** of event v at device n.
- μ_v^n is the **baseline intensity** of event v at device n.
- $Nei(n)$ is the set of **neighboring devices** of device n which can be known from the topology matrix $\mathbf{B}$.
- $a_{v_i v}$ indicates the **activation effect** of event type v_i on event type v, which is assumed to follow the principle of **Granger Causality**.

[1]. Hawkes, Alan G, et al. "Spectra of some self-exciting and mutually exciting point processes." *Biometrika* 58.1 (1971).

[2]. Cai, Ruichu, et al. "THPs: Topological hawkes processes for learning causal structure on event sequences." IEEE Transactions on Neural Networks and Learning Systems (2022).

Challenges

- **Scalability Challenge**:

The scales of the problems presented in this competition ranges from tens to a hundred, which is considered a significant hurdle for causal discovery. Finding an **efficient solution** to problems of such scale is a daunting task.

- **Effectiveness Challenge**:

The TNFD task is closely related to the livelihood infrastructure, incorrect outcomes could lead to severe economic losses and negative social public opinion. As a result, it presents a challenge to the **accuracy** of causal discovery.

- **Interpretability Challenge**:

In order to obtain results that are comprehensible to humans, it is imperative that the discovered causal graph be a **directed acyclic graph (DAG)**. However, ensuring this constraint satisfied during the optimization process poses a challenge.

To address the above challenges, we propose **S²GCSL: a Simple yet Scalable Granger Causal Structural Learning Approach** for **fast and effective** causal discovery.

Event Sequence $\boldsymbol{X} = \{(v_i, t_i, n_i) \mid i = 1, \ldots, m\}$

Maximum Likelihood Estimation

$$\lambda_v^n(t) = \mu_v^n + \sum_{i: n_i \in Nei(n), t_i < t} a_{v_i v} \phi(t - t_i)$$

$$\boldsymbol{A}'' = [a_{v_i v}] \in \mathbb{R}^{|V| \times |V|}, \boldsymbol{\mu} \in \mathbb{R}^{|V|}$$

- $\boldsymbol{A}''$ and $\boldsymbol{\mu}$ are to-be-estimated parameters.

$$L(A'', \mu) = \sum_n \left(\sum_{i=1}^{m_n} \log \lambda_{v_i}^n(t_i) - \sum_{v=1}^{V} \int_0^T \lambda_v^n(t) dt \right) \longrightarrow \boldsymbol{A}''_\star = \arg\min_{\boldsymbol{A}'', \boldsymbol{\mu}} -L(\boldsymbol{A}'', \boldsymbol{\mu}).$$

- Convert the causal discovery problem into an optimization problem

Constrained Gradient Descent based Maximum Likelihood Estimation

$$\boldsymbol{A}''_{\star} = \arg\min_{\boldsymbol{A}'',\boldsymbol{\mu}} - L(\boldsymbol{A}'', \boldsymbol{\mu}).$$

- For **Scalability Challenge**: Gradient descent
- For **Effectiveness Challenge**: Entry-norm Penalty $\|\boldsymbol{A}''\|_{1,1}$
- For **Interpretability Challenge**: Acyclic Constraint [1] $h(\boldsymbol{A}'') = \mathrm{trace}\left[(\boldsymbol{I} + \boldsymbol{\alpha}\boldsymbol{A}'' \circ \boldsymbol{A}'')^{|\boldsymbol{V}|}\right] - |\boldsymbol{V}|$

Final Objective: $\boldsymbol{A}''_{\star} = \operatorname{argmin}_{\boldsymbol{A}'',\boldsymbol{\mu}} -L(\boldsymbol{A}'', \boldsymbol{\mu}) + \lambda_1 \|\boldsymbol{A}''\|_{1,1} + \lambda_2 h(\boldsymbol{A}'')$

Optimization

We employ the **Adam** [2] optimizer to solve the above problem.

Pruning

After the above process converges, we will delete edges that are below a predefined threshold to obtain the final causal graph

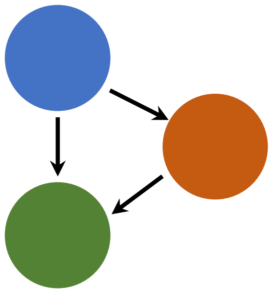
$$\boldsymbol{A}' = \boldsymbol{A}''_{\star} \geq \rho$$

[1]. Yue Yu, et al. "DAG-GNN: DAG Structure Learning with Graph Neural Networks." Proceedings of the 36th International Conference on Machine Learning (2019)

[2]. Kingma, et al. "Adam: A Method for Stochastic Optimization." Proceedings of the 3rd International Conference for Learning Representations (2014).

Experiment: Setup

Simulation:

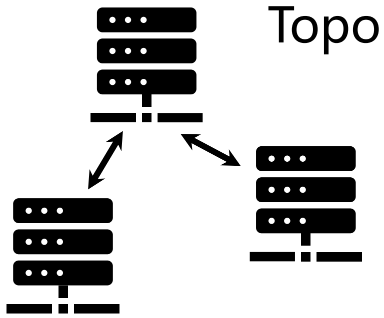


Random DAG

Topological Hawkes Process

Simulated Event Sequences

$$\boldsymbol{X} = \{(v_i, t_i, n_i) \mid i = 1, \ldots, m\}$$



Random Device Topology

Simulation parameters:

- Alarm types $(|N|)$: $\{20, 40, 60, 80\}$
- Devices $(|V|)$: $\{5, 10, 15, 20, 25, 50, 100\}$
- Sample size (m) : $\{50k, 100k, 150k, 200k, 250k, 300k\}$
- μ range$(\times 10^{-5})$: $\{(1,3), (3,5), (5,7), (7,9)\}$
- α range$(\times 10^{-5})$: $\{(1,2), (2,3), (3,4), (4,5), (5,6)\}$
- Time interval Δ: $\{(1,2), (2,3), (3,4), (4,5), (5,6)\}$

Experiment: Results on Simulation datasets

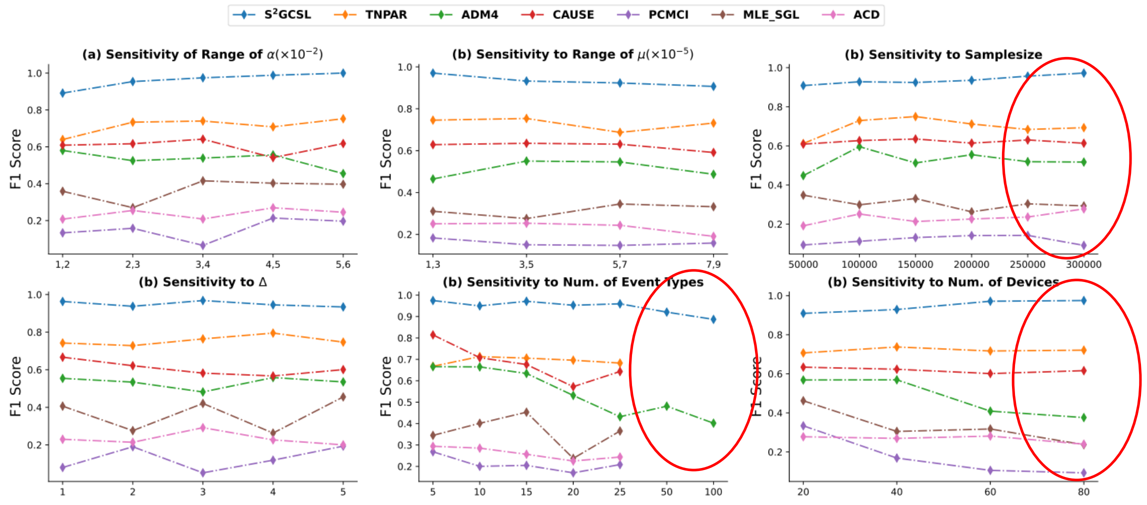


Figure 2: The F1 Scores of different methods on synthetic data.

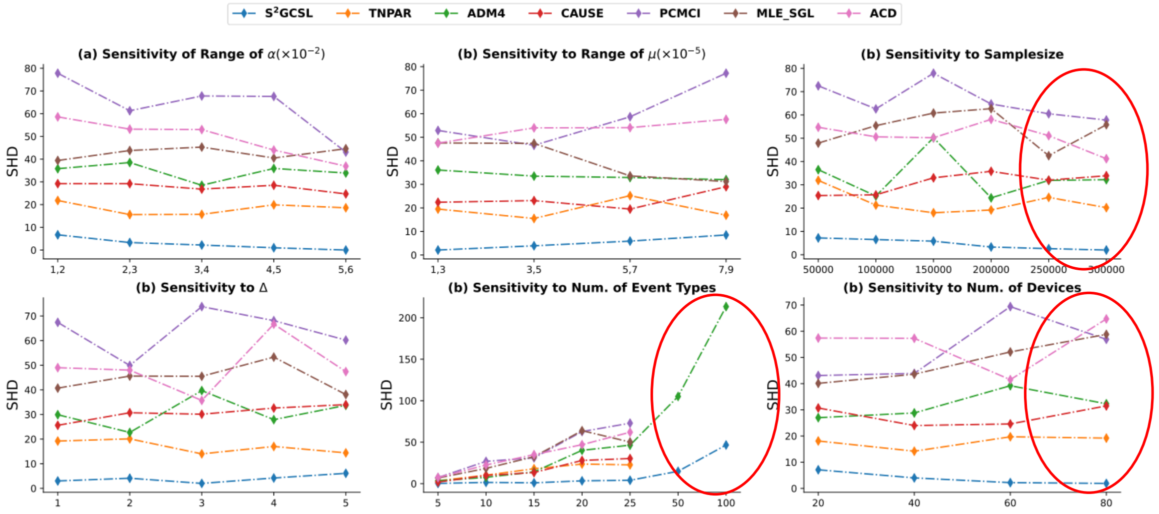


Figure 3: The SHD of different methods on synthetic data.

Metrics:

- F1 Score ↑
- Structural Hamming Distance (SHD) ↓
- Structural Interventional Distance (SID) ↓
- Wall-clock Execution Time (ET) ↓

S^2GCSL surpass all the other compared algorithms on effectiveness (F1 Score, SHD and SID), especially on **large-scale** problems

The larger the problem scale, the more pronounced the advantages for S^2GCSL.

Experiment: Results on Simulation datasets

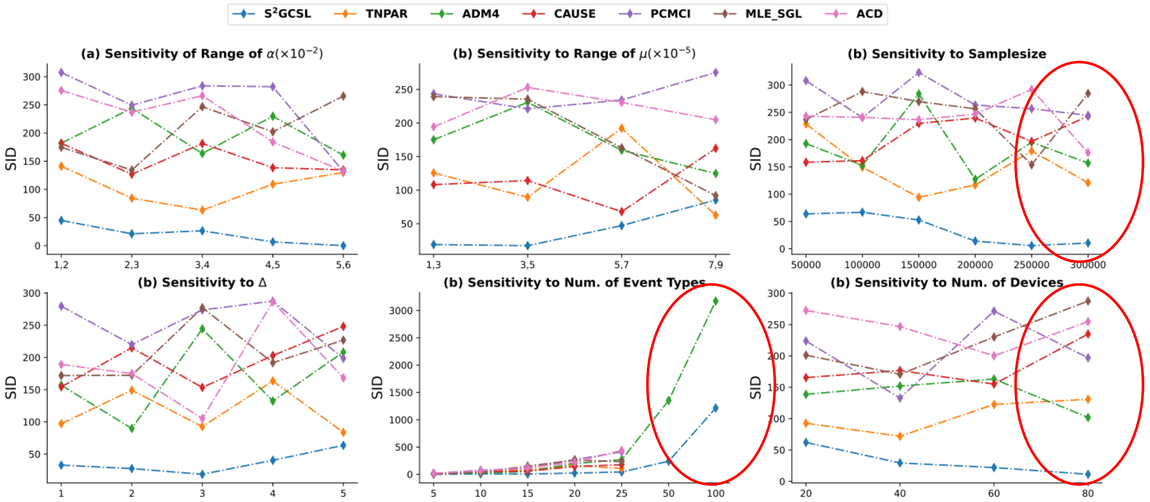


Figure 4: The SID of different methods on synthetic data.

S^2GCSL surpass all the other compared algorithms on effectiveness (F1 Score, SHD and SID), especially on **large-scale** problems

The larger the problem scale, the more pronounced the advantages for S^2GCSL.

Table 1: The wall-clock execution time (s) of different methods on different scale of synthetic problems. The algorithm with the highest efficiency under each scale of problem is marked in bold, and "-" indicates that results cannot be obtained within one hour.

Algorithms	5	10	15	20	25	50	100
S^2GCSL	$\mathbf{2.48 \times 10^0}$	$\mathbf{8.67 \times 10^0}$	$\mathbf{1.92 \times 10^1}$	5.48×10^1	8.64×10^1	$\mathbf{2.11 \times 10^2}$	$\mathbf{5.42 \times 10^2}$
TNPAR	3.61×10^2	6.40×10^2	7.82×10^2	9.93×10^2	1.46×10^3	-	-
ADM4	1.17×10^1	2.11×10^1	3.00×10^1	$\mathbf{4.46 \times 10^1}$	$\mathbf{6.68 \times 10^1}$	2.51×10^2	7.58×10^2
CAUSE	6.88×10^2	9.05×10^2	1.21×10^3	1.66×10^3	1.92×10^3	-	-
PCMCI	1.70×10^1	2.58×10^2	8.91×10^2	1.78×10^3	2.86×10^3	-	-
MLE_SGL	1.28×10^2	3.22×10^2	6.04×10^2	8.23×10^2	1.08×10^3	-	-
ACD	3.80×10^1	9.90×10^1	1.93×10^2	2.35×10^2	4.70×10^2	-	-

S^2GCSL remains competitive or surpasses other compared algorithms in efficiency across problems scale ranging from 5 to 100

Up to 277x acceleration!

Experiment: Results on Real-world datasets

Real-world Metropolitan Telecommunication Network Alarm Data

Table 2: Performances of different methods on metropolitan telecommunication network alarm data. The algorithm perform best under each metric is highlighted in bold.

Algorithms	F1 Score (↑)	SHD (↓)	SID (↓)	ET(s)(↓)
S^2GCSL	$\mathbf{0.40_{\pm 0.06}}$	$\mathbf{60.6_{\pm 6.59}}$	$397_{\pm 30.2}$	**737**s
TNPAR	$0.23_{\pm 0.06}$	$83.1_{\pm 6.07}$	$543_{\pm 62.6}$	4604s
ADM4	$0.19_{\pm 0.03}$	$83.5_{\pm 4.15}$	$475_{\pm 50.4}$	861s
CAUSE	$0.29_{\pm 0.04}$	$78.1_{\pm 4.09}$	$468.7_{\pm 29.4}$	7209s
PCMCI	$0.08_{\pm 0.02}$	$75.5_{\pm 4.32}$	$\mathbf{367_{\pm 17.1}}$	9342s
MLE_SGL	$0.19_{\pm 0.05}$	$77.2_{\pm 4.77}$	$406_{\pm 29.0}$	3253s
ACD	$0.14_{\pm 0.04}$	$107_{\pm 6.07}$	$655_{\pm 54.6}$	1943s

Most promising in real-world scenarios

S^2GCSL surpasses other compared algorithms in F1 Score, SHD and ET on real-world dataset

Conclusion & Take-home Message

- **Effective and Scalable Solution**: S^2GCSL introduces an effective and scalable approach for Granger causal structural learning from topological event sequences, optimized for large-scale telecommunication network fault diagnosis.

- **Key Methodology**: Linear kernel with gradient descent optimization; Incorporate expert knowledge via constraints to ensure interpretability.

- **Performance Advantage**: Demonstrates superior effectiveness and scalability on synthetic and real-world datasets compared to existing methods.

- **Practical Impact**: Addressing real-world fault diagnosis challenges through efficient Granger causal structure learning.

華東師範大學 EAST CHINA NORMAL UNIVERSITY

华东师范大学计算机科学与技术学院 School of Computer Science and Technology

NEURAL INFORMATION PROCESSING SYSTEMS

THANK YOU !




Mingjia Li* | Shuo Liu* | Assoc. Prof. Hong Qian† | Prof. Aimin Zhou

(* Equal contribution † Corresponding author)

Source Code

Paper

Wechat

Shanghai Institute of AI for Education
School of Computer Science and Technology
East China Normal University, China

NeurIPS 2024

Email: limj@stu.ecnu.edu.cn